CEF2 RailDataFactory

D2.3 – High-speed pan-European Railway Data Factory Backbone Network

Due date of deliverable: 30/06/2023

Actual submission date: 12/07/2023

Resubmission date: 11/08/2023

Leader/Responsible of this Deliverable: Julian Wissmann (WP 2 lead)

Reviewed: Y

Document status		
Revision	Date	Description
01	09/03/2023	Document template generated
02	07/06/2023	Content transferred from Confluence
03	08/06/2023	First draft complete
04	18/06/2023	Clean version for advisory board review
05	29/06/2023	Final version after addressing of all advisory board comments
06	12/07/2023	Version submitted to project officer
07	11/08/2023	Disclaimer updated based on the feedback of the granting authority

Project funded by the European Health and Digital Executive Agency, HADEA, under Connecting Europe Facilities Digital Grant Agreement 101095272		
Dissemination Level		
PU	Public	X
SEN	Sensitiv – limited under the conditions of the Grant Agreement	

Start date: 01/01/2023 Duration: 9 months

ACKNOWLEDGEMENTS

This project has received funding from the European Health and Digital Executive Agency, HADEA, under Connecting Europe Facilities Digital Grant Agreement 101095272.

REPORT CONTRIBUTORS

Name	Company
Alexander Heine	DB
Jens Dalitz	DB
Julian Wissmann	DB
Wolfgang Albert	DB
Patrick Marsch (only editorial)	DB

Note of Thanks

We would like to thank our Advisory Board Members Maria Aguado, Saro Thiyagarajan, Oliver Lehmann and Manuel Kolly for the valuable discussion and in particular Xiaolu Rao and Janneke Tax for their thorough reviews of this deliverable and input to this work! Also thanks to Mayutan Arumaithurai for his review!

Disclaimer

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them.

Furthermore, the information in this document is provided “as is”, and no guarantee or warranty is given that the information is fit for any particular purpose. The author(s) and project consortium do not take any responsibility for any use of the information contained in this deliverable. The users use the information at their sole risk and liability.

Licensing

This work is licensed under the dual licensing Terms EUPL 1.2 (Commission Implementing Decision (EU) 2017/863 of 18 May 2017) and the terms and condition of the Attributions- ShareAlike 3.0 Unported license or its national version (in particular CC-BY-SA 3.0 DE).

EXECUTIVE SUMMARY

The European rail sector is currently on the verge to the strongest technology leap in its history, with many railway infrastructure managers and railway undertakings striving toward large degrees of automation in rail operation, and mechanisms to increase the capacity and quality of rail operation.

In particular in the pursuit of fully automated driving (so-called Grade of Automation 4, GoA4), where sensors and cameras on trains will be used to automatically react to hazards in rail operation, it is commonly understood that an individual railway company or railway vendor would not be able to collect enough sensor data to sufficiently train the artificial intelligence (AI) eventually deployed in the rail system.

For this reason, it is commonly assumed that a form of pan-European Rail Data Factory is needed, as an infrastructure and ecosystem that allows various railway players and suppliers to collect and process sensor data, perform simulations, develop AI models, certify models, and ultimately deploy the models in the automated railway system.

This document therefore describes how new and existing data centers and data sources can work together. It becomes clear what is required to create an appropriate infrastructure, which measures must be taken into account and what options there are for this complex infrastructure. Finally, how data exchange is carried out securely across countries and which options are available for this.

Also included are already existing networks that can be used to avoid increased costs and expenses in this complex construct, as far as possible.

An overview of the possible and required network security zones, which are required for the creation of this ecosystem, complete this overall picture.

ABBREVIATIONS AND ACRONYMS

Abbreviation	Definition
AI	Artificial Intelligence
bbIP	Railway operational IP network (German: „bahnbetriebliches IP Netzwerk“)
DSTW	Digital Interlocking (Digitales Stellwerk)
DWDM	Dense Wavelength Division Multiplexing
ETCS	European Train Control System
GoA4	Grade of Automation 4
HADEA	European Health and Digital Executive Agency
IAM	Identity and Access Management
IM	Infrastructure Manager
ISMS	Information Security Management Systems
ISP	Internet Service Provider
QoS	Quality of Service
RAMS	Reliability, availability, maintainability and safety
RU	Railway Undertaking
Rail-IX	Rail Internet Exchange for peering data factories
SD-WAN	Software-Defined WAN

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

The European railway sector is on the verge to the strongest technology leap in its history, with many railway infrastructure managers (IMs) and railway undertakings (RUs) striving toward large degrees of automation in rail operation, and mechanisms to increase the capacity and quality of rail operation.

In particular, various railway companies - both IMs and RUs - and railway suppliers are currently working toward fully automated rail operation (so-called Grade of Automation 4, GoA4), for instance in the context of the Shift2Rail [1] and Europe's Rail [2] programs, in which sophisticated lidar and radar sensors as well as cameras are used to automatically detect and respond to hazards in rail operation, such as objects on the track or passengers in stations in dangerous proximity of the track. Another important use case is high-precision train localisation by detecting static infrastructure elements and locating them on a digital map, as for instance covered in the Sensors4Rail project [3]. While the rail system has various properties that render fully automated driving principally easier than, e.g., in the automotive sector (for instance, railway motion is only one-dimensional, scenarios are typically much less complex than automotive scenarios, etc.), key challenges on the way to fully automated driving in the rail sector are that hazardous situations have to be detected much earlier due to long braking distances, and it is very challenging to collect and annotate sufficient amounts of sensor data with sufficient occurrences of relevant incidences to perform the required artificial intelligence (AI) training and to be able to prove that the trained AI meets the safety and security needs.

For this, it is expected that single railway suppliers, IMs and RUs will not be able by themselves to collect and annotate sufficient amounts of sensor data for AI training purposes - but instead, a European data platform and ecosystem is required into which railway stakeholders (suppliers, IMs, RUs, railway undertakings, safety authorities, and others) can feed, process and extract sensor data, as well as simulate artificial sensor data, and through which the stakeholders can jointly develop and assess the AI models needed for fully automated driving.

1.1 AIM AND SCOPE OF THE CEF2 RAILDATAFACTORY STUDY

The CEF2 RailDataFactory study focuses exactly on aforementioned vision of a pan-European Rail Data Factory for the joint development of fully automated driving. The study, being co-funded through HADEA, aims to assess the feasibility of a pan-European Rail Data Factory from technical, economical, legal, regulatory and operational perspectives, and determine key aspects that are required to make a pan-European Rail Data Factory a success. For a better understanding of the study's aim and scope, please see Chapter 1.1 in Deliverable 1 [4].

1.2 DELINEATION FROM AND RELATION TO OTHER WORKS

The notion of Rail Data Factory is also covered in other work, such as the Shift2Rail project TAURO [5] or Europe's Rail Innovation Pillar FP2 R2DATO project [6]. Further, Deutsche Bahn, within the sector initiative "Digitale Schiene Deutschland", has already started setting up some related data center components [7]. For a better understanding of the relationship between the CEF2 RailDataFactory study and these works, please see Chapter 1.2 in Deliverable 1 [4].

1.3 AIM AND STRUCTURE OF THIS DELIVERABLE

This current document is the deliverable D 2.3 of the CEF 2 RailDataFactory project, covering the High-speed pan-European Railway Data Factory Backbone Network and specifically aiming at providing requirements and implementation proposals for the same.

The aim of the document is to obtain early feedback and possible additions from the sector on the architecture and building blocks, in order to update the work accordingly and consider the obtained input in the subsequent phases of the project, in which the operational, legal, regulatory and business aspects related to the pan-European Data Factory will be investigated, and possible deployment scenarios will be explored.

The remainder of this document is structured as follows:

- In Chapter 2, the High-speed pan-European Railway Data Factory Backbone Network is introduced;
- In Chapter 3, network implementation options are presented;
- In Chapter 4, proposals are provided regarding how national operational networks could be connected;
- And finally, in Chapter 5, a summary is provided.

2 HIGH-SPEED PAN-EUROPEAN RAILWAY DATA FACTORY BACKBONE NETWORK

The idea of a High-speed Pan-European Railway Data Factory Backbone Network is to enable the exchange of data across the European Data Factories and consequently enable machine learning in such a way that all required data, which is not marked as private by the data owner, from all data sources can be made available securely and efficiently.

The following architecture guidelines enforce compliance and possibilities for a pan-European Railway Data Factory Network. The network enables the benefit for all participants to work together across all European countries.

By choosing an approach that follows principles of loose coupling and federation, the resulting distributed system will avoid the building of silos and isolated solutions. Also, this allows to leverage the market to create new ideas and build advanced smart services in industries and innovations enabled by pan-European data.

There are several approaches to making the idea of allowing everyone to access data that is to be shared. These can be seen in the following diagrams and are also described in more detail below.

In general, an Identity and Access Management (IAM) will take care of the access and as well of the trust management for all interactions within the pan-European Data Factory. Multi-tenancy will take care of the sovereignty of data and data exchange.

2.1 GENERAL REQUIREMENTS FOR THE HIGH-SPEED PAN-EUROPEAN RAILWAY DATA FACTORY BACKBONE NETWORK

For the design of a High-speed pan-European Railway Data Factory Backbone Network for the exchange of large amounts of data between manufacturers, railway infrastructure companies and railway undertakings, several important factors should be considered. Below are some key points and recommendations to consider when planning and designing such a network:

1. **Robust and fast connections**: Invest in fiber or lease so-called dark fibre to ensure maximum speed and reliability. Fibre offers enormous bandwidth, which is ideal for transferring large amounts of data.
2. **Geographical distribution**: When designing the network, consider the locations of the parties involved. It must be ensured that all locations are connected by a high-speed link.
3. **Redundancy**: The network should be designed with redundancy to ensure that if one part of the network fails, it will still function. A ring design can help with this by providing alternative routes for data.
4. **Scalability**: The design should also consider the possibility for future growth. It should be planned beyond current needs, and allow for the possibility of adding more sites or higher volumes of data in the future.
5. **Security**: Since sensitive and potentially business-critical data is being handled, security is a must. Encryption of data in transit, strong authentication protocols and regular security audits are some of the ways to ensure this.

6. **Monitoring and maintenance**: Good network operations require constant monitoring and maintenance. Tools to monitor network traffic and identify bottlenecks or outages should be deployed. A dedicated rail backbone network IT team should be ready to respond to problems and maintain the network regularly.
7. **Connectivity to Cloud platforms**: Depending on specific requirements, connections to Cloud platforms such as AWS, Google Cloud or Microsoft Azure should be considered. These platforms can provide additional storage and computing resources and are ideal for providing data analytics services.
8. **Connection to the operational railway infrastructure networks**: In order to seamlessly transfer the data from the trains or the infrastructure to the Data Factory, a non-reactive connection to the internal networks of the railway infrastructure providers and the railway undertakings would be recommended.
9. **Functional safety**: A functionally safe system requires a reliable and secure backbone network to ensure that safety-critical data is transmitted correctly, unaltered and in a timely manner. In addition, the network must also be able to detect and defend against potential threats or attacks, record changes in an audit-proof manner to meet functional safety standards.

2.2 Data Collection Backbone

At Deutsche Bahn, a network for future railway operations is in the process of being setup, the so-called bbIP (“bahnbetriebliches IP Netzwerk”) network [8]. This network is owned and operated by Deutsche Bahn and its purpose is to be the underlying communications layer for future railway applications such as digital interlockings or the European Train Control System (ETCS). As such, it fulfils the highest integrity and availability demands. In order to digitalise their networks, other infrastructure operators are working on similar operational networks.

The data handled on these networks is typically part of the safety and operations critical domain of the rail network. As such they are not connected to networks outside of the infrastructure operators control. Therefore, direct transmission of data from these networks into third party networks is typically not permitted.

Besides such high security networks, there are typically also other operational networks present in railway operations, for example those for mobile communications backhaul used by GSM-R and in the future FRMCS [9].

With autonomous driving being a safety critical system, as well, it is expected that an operational GoA4 system would be required to use railway operational networks as well.

As these networks are the basis for the digitalisation of the railway system, they will need to reach into every corner of the railway network where an advanced digitalisation will be rolled out. Therefore, making them the basis for connecting up data sources in the field whose data is used in the development of a GoA4 system would be a natural choice, as these networks are specifically built to handle the safety and security requirements of such a data source. Availability and bandwidth can be guaranteed through Quality of Service (QoS) mechanisms. It is up to future studies of these networks to determine which kind of operational network could be leveraged to host a Data Factory data collection backbone.

However, when we look to the bandwidth requirements for a Data TouchPoint (enabling the data exchange from trains to ground, see [4]), which we assume to be in the order of tens of Gbps, these are not reflected in the capabilities of the infrastructure operators’ networks today. Since a data

TouchPoint will also require a certain amount of computation resources to preprocess data and additional elements to handle security aspects, like transport encryption, can be located there, a railway operator may also connect TouchPoints via public networks if the bandwidth requirements cannot be met on operational networks or other reasons speak for not using operational networks for this use case.

2.3 DATA FACTORY BACKBONE

Developing a dedicated backbone network for data transmission between railway companies can be a complex but rewarding task, especially when considering the need for high functional safety and the specific requirements of railway operations. This network can be specifically designed to support the development and operation of Grade of Automation 4 (GoA4) with the specific challenge of command-and-control infrastructure and vehicles.

An important aspect is the evaluation of specific railway standards, such as EN 50159 (“Safety-related communication in transmission systems”) and EN 50126 (“Railway Applications - The Specification and Demonstration of Reliability, Availability, Maintainability and Safety”). To meet the Data Centers functional safety requirements according to EN 50126, the backbone network may need to be designed to provide reliability, availability, maintainability, and safety (RAMS). The design and testing of functional safety must be carried out in accordance with the requirements of EN 50126, which covers the entire lifecycle of the system from conception, implementation, maintenance, and decommissioning.

Furthermore, the railway sector is classified as critical infrastructure as per the NIS2 directive (DIRECTIVE (EU) 2022/2555) [10]. As such, railway infrastructures are subject to strict regulation and special protection requirements. As this applies also to the networks, special requirements regarding protective measures and emergency management apply, which are taken into account within the ISO 27001 certification framework. There are several aspects to consider, one of which is ISO 27001 certification, an internationally recognised standard that provides a framework for Information Security Management Systems (ISMS). For a backbone network, this means that all aspects of data security, including physical security and access control, must be managed to high standards.

Operator responsibility is another important aspect in the construction and operation of the backbone network. This includes proper maintenance and upkeep of the network, implementation and compliance with security standards and policies, as well as providing sufficient capacity to handle data traffic. Moreover, the operator must respond to changes in network utilisation or the threat landscape, and proactively take measures to ensure network performance and security.

One strategy for setting up this network would be utilising existing dark fiber infrastructures, such as those offered by DB Netz, which can be both cost-effective and technically advantageous. Dark fiber provides high bandwidth, low latency, and high security, which are important for safety-critical applications like GoA4. Additionally, also Dense Wavelength Division Multiplexing (DWDM) could be used, as it greatly increases bandwidth, however coming at an additional cost. In the process of digitalising the railway network, infrastructure managers generally have a need to build up such fiber networks in order to connect their field elements to the control infrastructure.

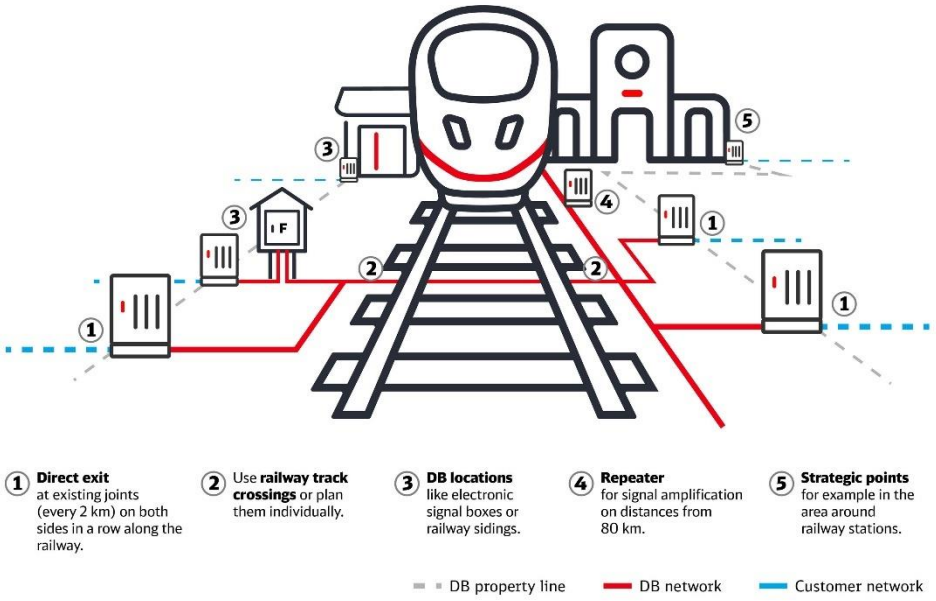

Figure 1: Dark fibre from DB Netz (DB Broadband)[1].

Another idea to approach this would be to set up one or more rail Internet exchange points built by infrastructure managers operating large amounts of fiber in the field. The rail Internet exchange would serve as the interconnection point, referred to as Rail-IX, for rail networks and the connection to Internet Service Providers (ISPs) in order to allow interconnection with public networks via which, e.g., industry partners can connect, and would involve several steps and components:

- **Site selection**: Firstly, a suitable location for the Rail-IX needs to be identified. Ideally, the site should be in proximity to data centers to ensure good connectivity.
- **Infrastructure**: Following the site selection, the next step is to build the necessary infrastructure. This includes racks, switches, fiber optic cables, power supply, and cooling systems. Providing a reliable and scalable infrastructure is crucial for efficiently handling network traffic.
- **Internet Service Providers (ISPs)**: To successfully operate the Rail-IX, ISPs that are interested in peering their networks at the Rail-IX need to be onboarded. This involves physically connecting their networks to the Rail-IX to enable the exchange of internet traffic.
- **Peering agreements**: Peering agreements must be established with the ISPs to define the terms and conditions for data traffic exchange. These agreements cover aspects such as capacity, routing policies, and costs.
- **Interconnection with other Rail-IX nodes**: To connect the proprietary Rail-IX with the larger commercial Internet exchange network, physical connections with other Rail-IX nodes need to be established. This enables the exchange of traffic with ISPs and networks connected at other Rail-IX locations.

[1] https://broadband.dbnetze.com/dbbroadband-en/Dark-Fiber

- **Operations and monitoring**: Once the Rail-IX is established, it needs to be operated and monitored. This includes monitoring network traffic, ensuring service quality, troubleshooting issues, and performing maintenance tasks to ensure smooth operation of the Rail-IX.
- **Governance**: Access to such a Rail-IX needs to be conducted in a fair and equal manner so that no partner is discriminated upon. This requires a clear governance framework for the operation of such a system across the entire EU.

It is important to note that building a proprietary Rail-IX infrastructure would be a complex task that would require significant investments in infrastructure, resources, and expertise. However, it may be a worthwhile investment as it strengthens data sovereignty in Europe and increases redundancy for new backbone networks by leveraging existing resources such as the existing dark fiber in the rail network.

There are several options when building backbone networks, including using existing dark fiber possessed by infrastructure providers, contracting service providers or building a new network connection. The choice between these options depends on various factors, including the specific requirements of the critical infrastructure.

In the following, a first comparison of the three solutions identified so far is presented, which in turn should only serve as an impulse for further investigations.

1. **Utilisation of existing dark fibers**: Utilising one's own, unused fiber-optic connections (dark fiber) can be advantageous in terms of cost and control. The entity has full control over its network infrastructure, offering greater flexibility in terms of usage and management. In addition, costs could be lower as there would be no need to pay monthly or annual fees to a third party. However, some disadvantages need to be taken into account. For instance, the entity must have the knowledge and resources to maintain and manage the network. Furthermore, issues might be encountered if the dark fiber isn't sufficient to meet the requirements or if upgrades or expansions are necessary.
2. **Use of service providers**: Using service providers can be a good option if the entity does not have the resources or knowledge to manage its own network. These companies usually offer a range of services, including network management, maintenance, and support. In addition, they can scale or adjust the network connection as needed. However, this option can be more expensive as the entity would have to pay for these services. Also, it may not have the same level of control over the network as when using its own dark fibers.
3. **New network connection construction**: Building a new network connection might be necessary if neither existing dark fibers nor the services of third parties meet the entity's requirements. This could be the case if there are specific demands regarding security, speed, or reliability. This option, however, requires substantial investments in terms of time and money. Also, the entity would need the necessary expertise to plan and implement such a project.

The advantages and disadvantages of the presented options are further detailed in Table 1.

Table 1: Comparison of backbone implementation options.

Option	Costs	Time	Flexibility	Security	Safety	Reliability
Use of existing dark fibers	No additional procurement costs as these resources are already in place. Maintenance and repair costs to equipment used to access the fibers may apply.	The existing infrastructure allows a faster network setup.	Full control and flexibility over the network.	Control over the network's security is in the hands of the owner of the dark fibers.	Safety is dependent on the initial installation and the ongoing integrity of the physical infrastructure	Reliability depends on the quality and condition of the existing dark fibers
Use of service providers	Costs can be high but might be cheaper compared to building a new network.	Quick network setup due to the provider's infrastructure and expertise.	Flexibility might be limited due to dependence on the provider's service.	Security can be very good, but control is less as it depends on the service provider.	Safety will depend on the provider's standards and protocols for their network management and operations.	High reliability should be offered by professional providers, but downtime that's beyond control may occur.
New network construction	High cost due to the materials and labor required.	The construction of a new network may take some time.	Full control and flexibility over the design of the network.	Full control over the network's security.	Safety depends on the construction standards, materials used, and the maintenance of the physical infrastructure.	High reliability can be achieved by ensuring the network complies with the latest standards.

For critical infrastructures, ensuring the reliability and security of the network is particularly important. Therefore, these aspects should be prioritised when deciding on one of these options. Using own dark fibers could provide more control and potentially more security and deterministic safety, while hiring service providers and building new network connections might offer more flexibility and scalability.

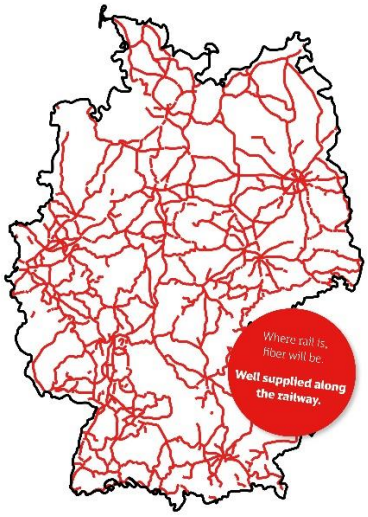

Figure 2: Example of 22.000 km existing fibre optic cables (33.400 km in 2026/2027)[2] in Germany.

In Germany, requirements related to operator responsibility according to the General Railway Act (AEG “Allgemeine Eisenbahngesetz” in Germany[3]) must also be considered. These relate to the safety, reliability, and sustainability of railway operations and set strict standards for railway infrastructure operators. Similar legal requirements may apply to other infrastructure operators in Europe, however this was not examined at this point.

In the context of Gaia-X [11] and Catena-X [12], two initiatives aimed at creating a secure and trustworthy data infrastructure in Europe, the backbone network should be designed for compatibility and interoperability. Gaia-X, the European initiative for a secure and transparent data infrastructure, can serve as a model for the backbone network. By applying Gaia-X principles, the network can ensure high data sovereignty, interoperability, and transparency. The experiences of Catena-X, an automotive network based on the principles of Gaia-X, can also be very useful. They can provide insights into best practices for building and operating a secure and reliable backbone network.

Although developing a dedicated backbone network may not be cost-effective compared to existing providers, leveraging existing infrastructure and the ability to meet specific requirements and standards of railway operations can bring significant benefits. It also offers greater control and adaptability, which can be crucial in safety-critical applications. With the right resources and proper planning, such a network can provide substantial advantages for the development and operation of GoA4.

[2] https://broadband.dbnetze.com/dbbroadband-en

[3] https://www.gesetze-im-internet.de/aeg_1994/

3 NETWORK IMPLEMENTATION OPTIONS

In the context of building a pan-European Railway Data Factory, three different approaches from the network perspective can be considered: a centralised approach, a federated approach, and a hierarchical approach. Each approach presents its own advantages and challenges in terms of data management, governance, and flexibility.

The centralised approach involves all users working together in a single pan-European Railway Data Factory. Functions like storage, compute resources, and identity access management are centralised. A high-speed backbone network connects data sources and ensures efficient data transfer. While this approach enables central data management, concerns arise regarding governance, ownership, and organisational challenges. It may also limit data management and support for new or experimental data types. The federated approach breaks down data silos and merges existing data centers into a common pan-European Railway Data Factory. Members can use their own data centers or connect to existing participants, with data catalogues ensuring data sovereignty and privacy. A pan-European backbone network facilitates data exchange, enabling connection to additional data sources. Members maintain data ownership and governance, following their own policies and frameworks. The hierarchical approach connects a network of data centers to a central storage repository. Relevant data is consolidated for the pan-European system, and data centers execute AI tasks. While centralisation enables efficient data sharing, duplication and distribution costs can arise. Only agreed-upon data with sufficient quality is stored centrally, but members have freedom to experiment with data in their own systems.

Each approach has trade-offs in data management, governance, and flexibility, which are further elaborated on in the following. The choice depends on the specific needs and considerations of the pan-European Railway Data Factory implementation.

3.1 CENTRALISED APPROACH

With the centralised approach, illustrated in Figure 3, all users of the member states work together in one pan-European Railway Data Factory. All functions provided by the Data Factory, from storage to compute resources and IAM would be centralised in this case.

A High-speed pan-European Railway Data Factory Backbone Network is connected to this pan-European Railway Data Factory to ensure that all desired and required data sources and TouchPoints can provide their data with the required bandwidths.

This approach enables central management of all data and newly arriving data, as well as their access. However, it raises concerns in the areas of data governance and data ownership as well as organisational questions which may be hard to overcome.

Additionally, this approach may pose limits in the areas of data management and supported data. If, e.g., a member wants to experiment with a new type of data that is not yet part of an agreed ontology, does not fit to an agreed data format or does not have the agreed upon quality, this approach may pose problems.

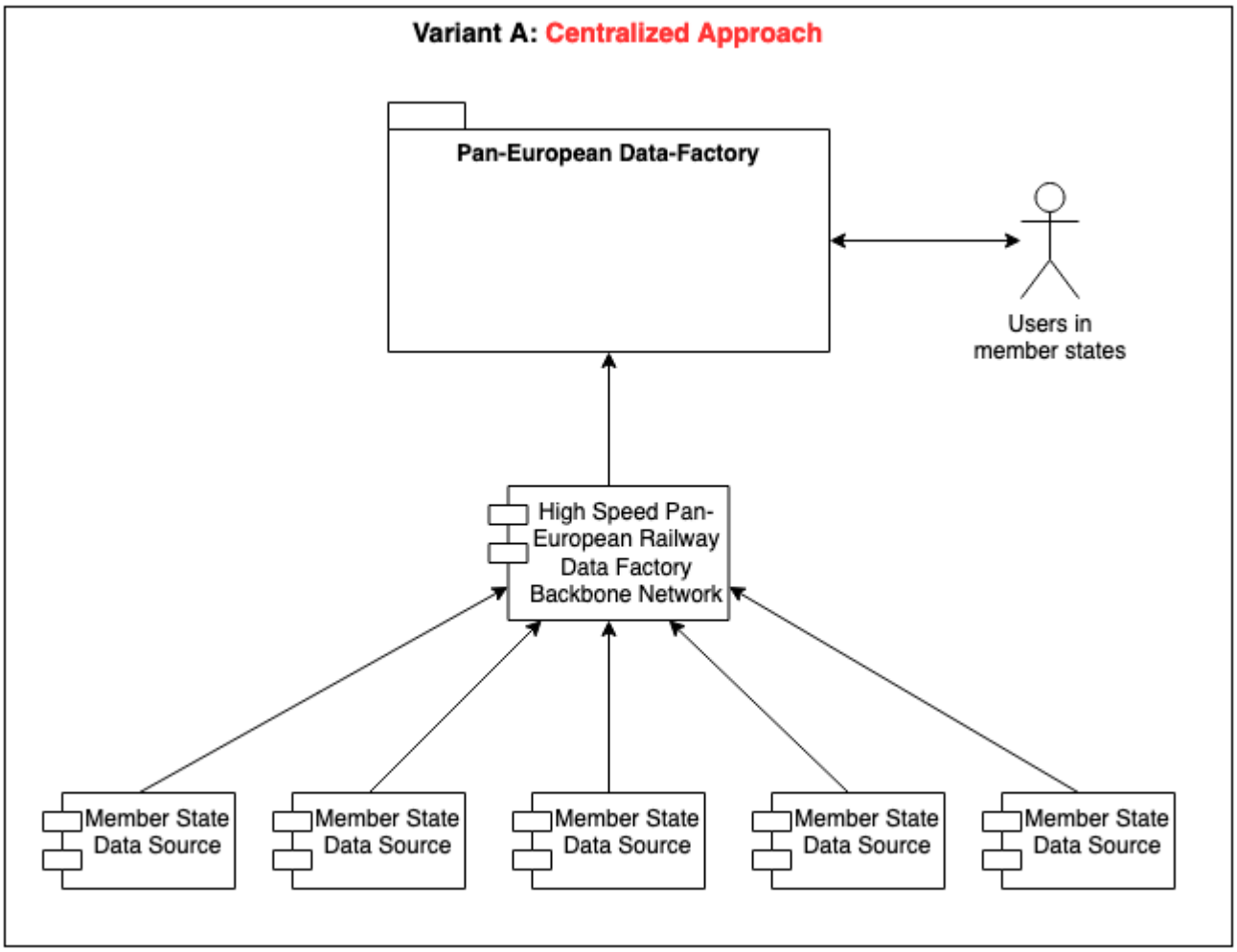

Figure 3: Centralised network implementation variant.

3.2 FEDERATED APPROACH

With the variant of a federated approach, as illustrated in Figure 4, it is possible to break up existing data silos and also to merge existing data centers and data sources into a common pan-European Railway Data Factory. This approach gives maximum flexibility. A European member can use their existing data center or connect to an existing participant of the grid and rent capacity there. Via a data catalogue, it is known for each data center what kind of data are generally available, though it is important to take care of data sovereignty and privacy.

To exchange data, a High-speed pan-European Railway Data Factory Backbone Networks comes in which includes components to provide the needed and expected bandwidth to transmit data in a traceable and efficient way. Now each user of each member state can get the data for their needs.

This way, additional data silos or data centers can be connected, which serve either for communication or for data delivery. This way, it doesn't matter whether the data comes directly from a train, from another data source or from an intermediate system such as a Data TouchPoint.

With this approach, data ownership and data governance is delegated to the members bringing in the data, thereby allowing every member to handle data according their own policies and frameworks.

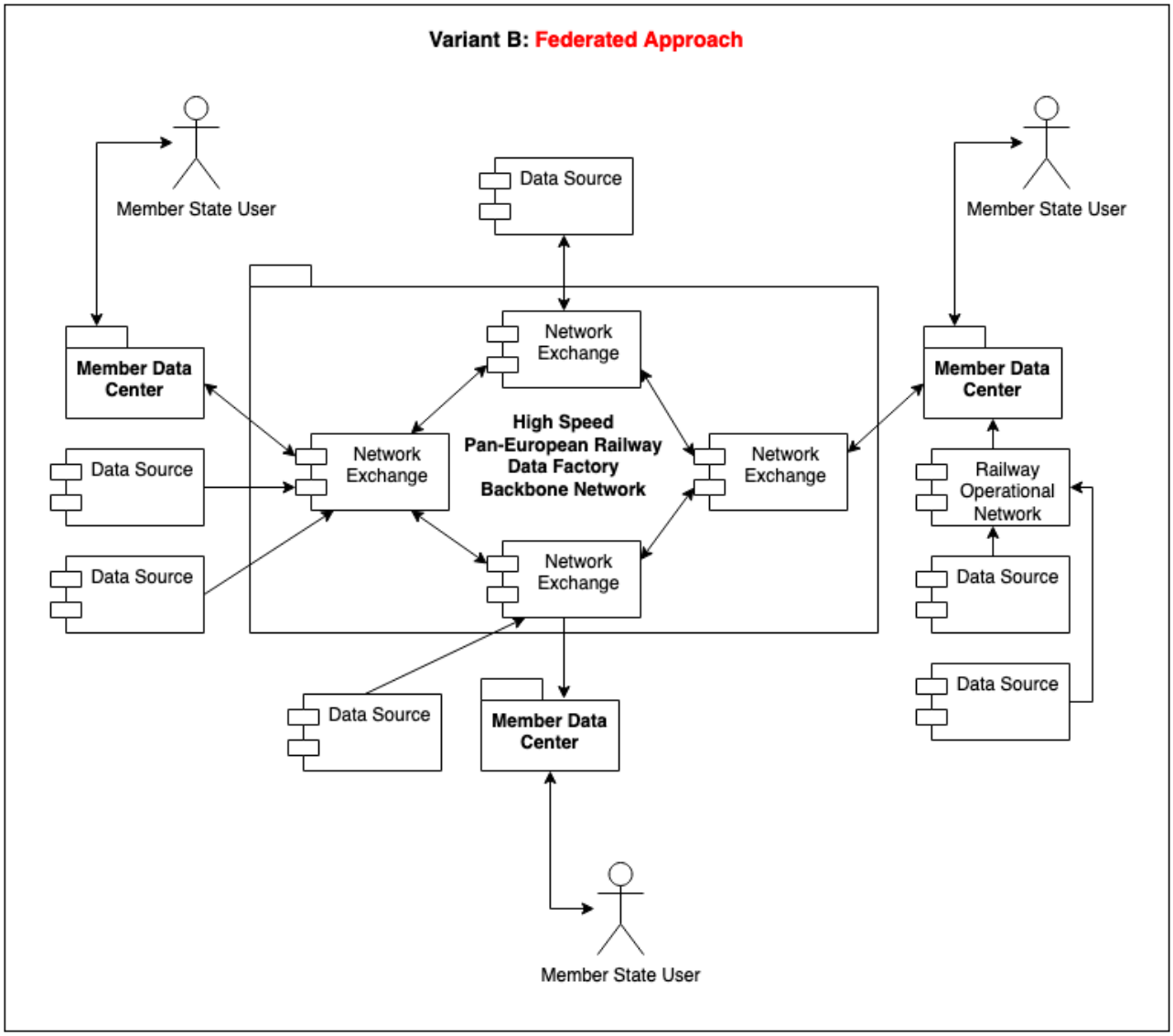

Figure 4: Federated network implementation variant.

3.3 HIERARCHICAL APPRAOCH

With a hierarchical approach, as shown in Figure 5, a network of data centers would be connected to a central storage which would consolidate all relevant data for the pan-European system. In this approach, members could bring in their own data center, though data would be held and distributed from a central repository.

The data would be stored in a central repository that is accessible to all data centers within the network. This central storage allows for efficient data sharing and access across the network. However, this notion of centralisation may cause significant data duplication and additional cost, as data would still need to be distributed to operators or industry partners' data centers for computation. It must be noted that in this approach, like in a centralised approach, only agreed-upon data with sufficient data quality would likely be suitable to be stored in the central repository. However, with organisations still bringing in their own data center, they would be free to experiment with such data in their own system.

This structure involves a central storage unit that houses the relevant data and models, while multiple data centers are responsible for executing the AI training tasks within the pan-European Data Factory.

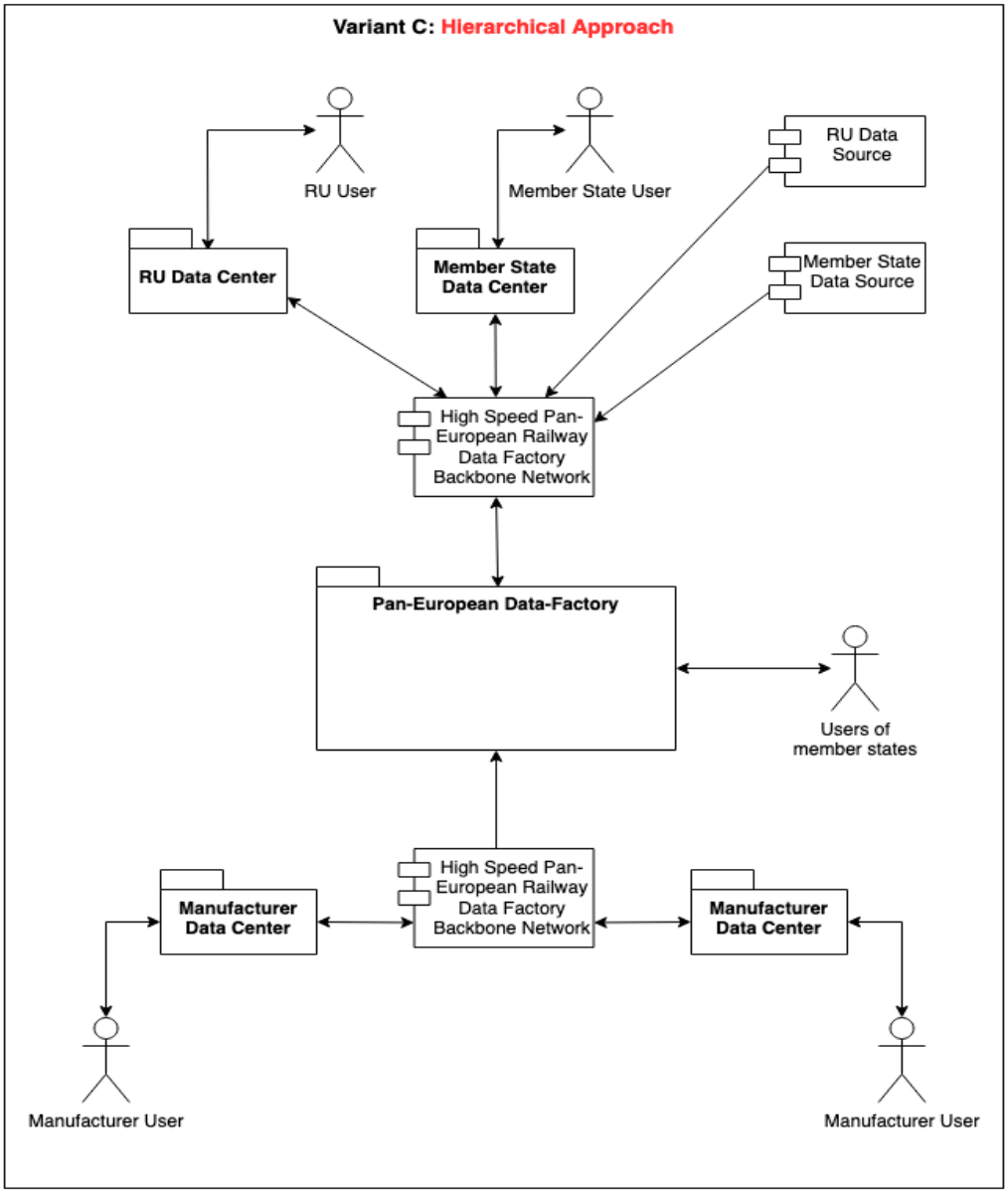

Figure 5: Hierarchical network implementation variant.

3.4 APPROACH COMPARISON

In Table 2, a comparison of the three network implementation variants is presented, and their specific characteristics are highlighted. In subsequent projects, an allocation to the national and EU-wide requirements must be made, taking into account the technical and non-technical requirements.

Table 2: Comparison of network implementation options.

Approach	Description	Advantages	Challenges
Variant A **Centralised Approach**	In the centralised approach, all users work together in a pan-European Railway Data Factory. The data factory provides centralised storage, compute resources, and IAM. A backbone network ensures data sources and touch points can transmit data effectively.	- Central management of all data and access - Central data governance and management	- Concerns regarding overall governance - Concerns regarding data governance and ownership - Organizational challenges - Limited support for new data types or formats
Variant B **Federated Approach**	The federated approach breaks data silos and allows merging of existing data centers and sources. Members can use their own data centers or connect to existing participants. A data catalogue provides information on available data. Data exchange occurs via a pan-European Railway Data Factory Backbone Network.	- Maximum flexibility and data sovereignty - Data centers can be connected without centralizing ownership - Efficient data transmission with traceability	- Need to ensure data sovereignty and privacy - Challenges in coordinating data centers and data sharing
Variant C **Hierarchical Approach**	The hierarchical approach connects a network of data centers to a central storage. Members can bring their own data centers, but data is held and distributed from a central repository. Data sharing and access are efficient across the network.	- Efficient data sharing and access - Centralised storage for consolidated data	- Concerns regarding overall governance - Concerns regarding data governance and ownership - Potential data duplication and increased cost - Distribution of data to operators or industry

			partners' data centers - Agreed upon data quality for storage in the central repository - Data experimentation limited to individual data centers - Challenges in executing AI training tasks across multiple data centers within the Data Factory

From the current point of view and with regard to security, data sovereignty and legal assessment, the federated approach as presented in Section 3.2 is the one that appears to be the most realistic for implementation, as it allows participants the flexibility to choose whether they want to integrate their own data centers or become a tenant in one as well as posing the least governance challenges as seen in Table 2.

The communication between the data centers should be handled in a separate network, with sufficient bandwidth, and apply zero trust concepts. As long as only a small number of data centers are involved, simpler security concepts and point-to-point connections may be used. However, the goal should be a network to connect any number of data centres without necessary trust relationships. Today, feasible networks can be rented from most international carriers. Apart from the secure operation of the network, security tasks (e.g., data centre firewall) must not be handled by the carrier, and the network should not be used for other tasks than the data transmission between the data centers (e.g., no maintenance or access for third parties). Both push (e.g., notifications) and pull connections (e.g., copy of new data) can be used between the data centers.

Networks with a wide range of safety and security requirements may be feasible for the data transmission between trains, data TouchPoints and data centers. Public networks and air-gapped networks are not considered in detail. Network types with regulations and requirements in between these extremes are shown by using the example of the railway network infrastructure of DB Netz (bbIP).

In general, these networks are internal networks of the infrastructure operator who operates various networks, from simple communication networks to networks for traffic control and management, with a wide variety of safety and security requirements. If a communication is established in between different networks, the traffic will be routed over a security component (e.g., a firewall), and the establishment of connections is only permitted in networks with the same or lower security level. Data may only be exported from networks with a high security level. External access to these

networks (e.g., for maintenance tasks) is only possible through individual approval processes, and permanent external connections are prohibited.

It can be assumed that data recorded in these networks will be forwarded through the infrastructure operators' network backbone, but can only be passed to networks with lower security levels. A data import or export by external access to these networks is usually prohibited by the security rules of the infrastructure operators. For this reason, a push procedure should be provided for the data transfer towards the data TouchPoint or data center.

4 INTERCONNECTING NATIONAL OPERATIONAL NETWORKS

In order to implement the interconnectivity with the High-speed pan-European Railway Data Factory Backbone Network, a common high-level understanding of network security zones is required. While security standards like ISO 62443 [13] go into much more detail, the following high-level analysis, as also illustrated in Figure 6, is assumed to provide enough context for the purpose of this proposal.

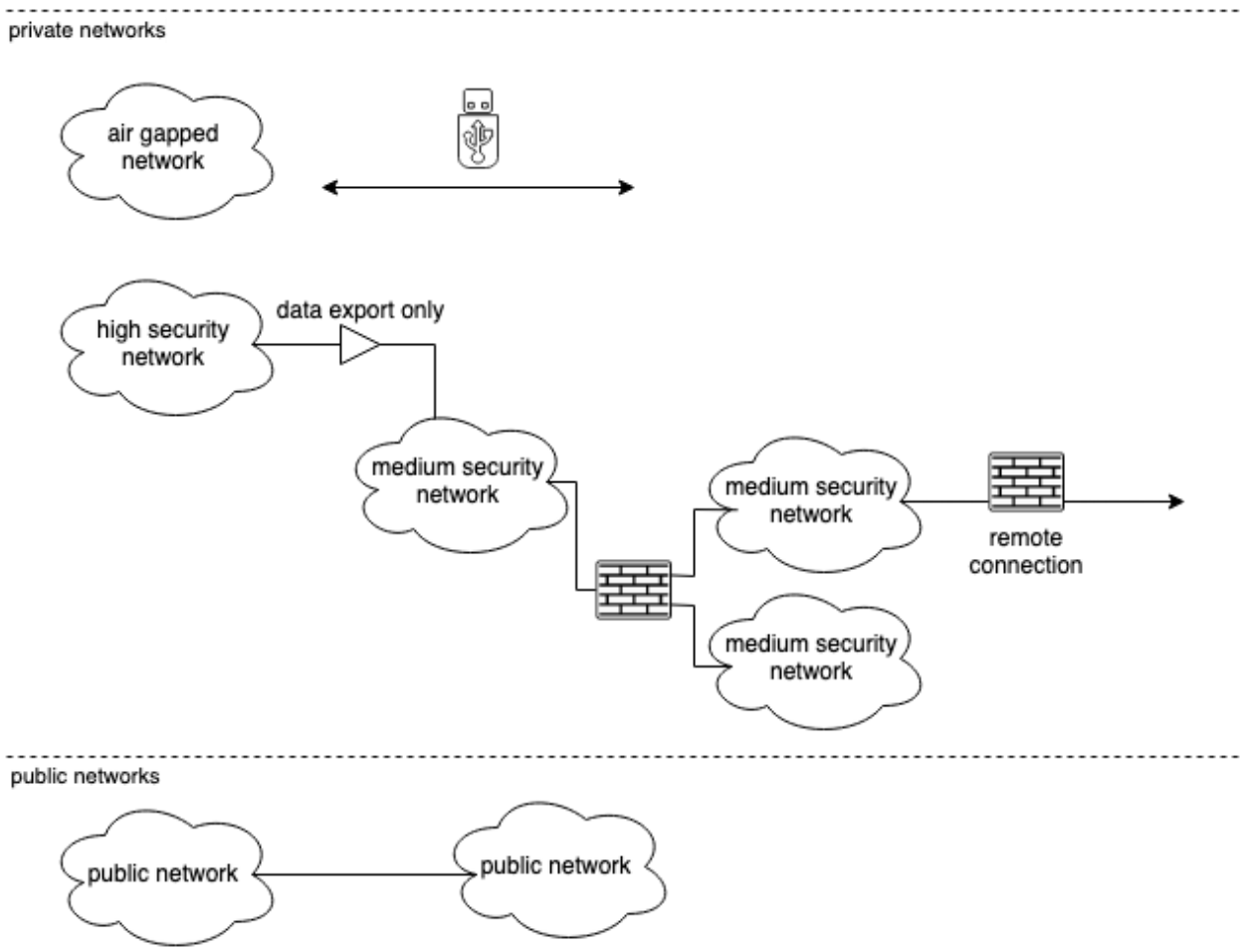

Figure 6: Network security zones.

Air Gapped Network

An air-gapped network refers to a computer network that is physically isolated from external networks, such as the internet or other interconnected networks. The term "air-gapped" derives from the concept that there is an "air gap" between the isolated network and the outside world, meaning there is no physical or electronic connection between the two.

High Security Network

A high security network is a computer network that incorporates stringent measures to protect sensitive data, critical systems, and communications from unauthorised access, cyber threats, and potential breaches. These networks are designed to provide the utmost level of security and typically involve advanced technologies, rigorous protocols, and robust defence mechanisms. Typically, these networks can communicate into networks of lower security classification, but not the other way around.

Medium Security Networks

A medium security network is a computer network that incorporates a moderate level of security measures to protect data, systems, and communications from unauthorised access and potential threats. While not as stringent as high-security networks, medium-security networks still employ various security measures to mitigate risks and ensure the confidentiality, integrity, and availability of network resources.

Public Networks

A public network refers to a computer network that is openly accessible to individuals and organisations, typically over the Internet or other shared communication infrastructure. Public networks are designed to facilitate connectivity and information exchange among users, often without requiring specific permissions or restrictions.

4.1 IMPLEMENTATION ON PRIVATE INTERCONNECTION

In order to interconnect an infrastructure manager's (IM) operational network, a site-to-site connection can be leveraged connecting the operational network to the data center using private networks, as shown in Figure 7. Security needs to be ensured on both ends by firewalls. This way, concerns can be separated clearly. It also enables to channel user access through existing networks on the IM side which allows the usage of existing security infrastructure in the network.

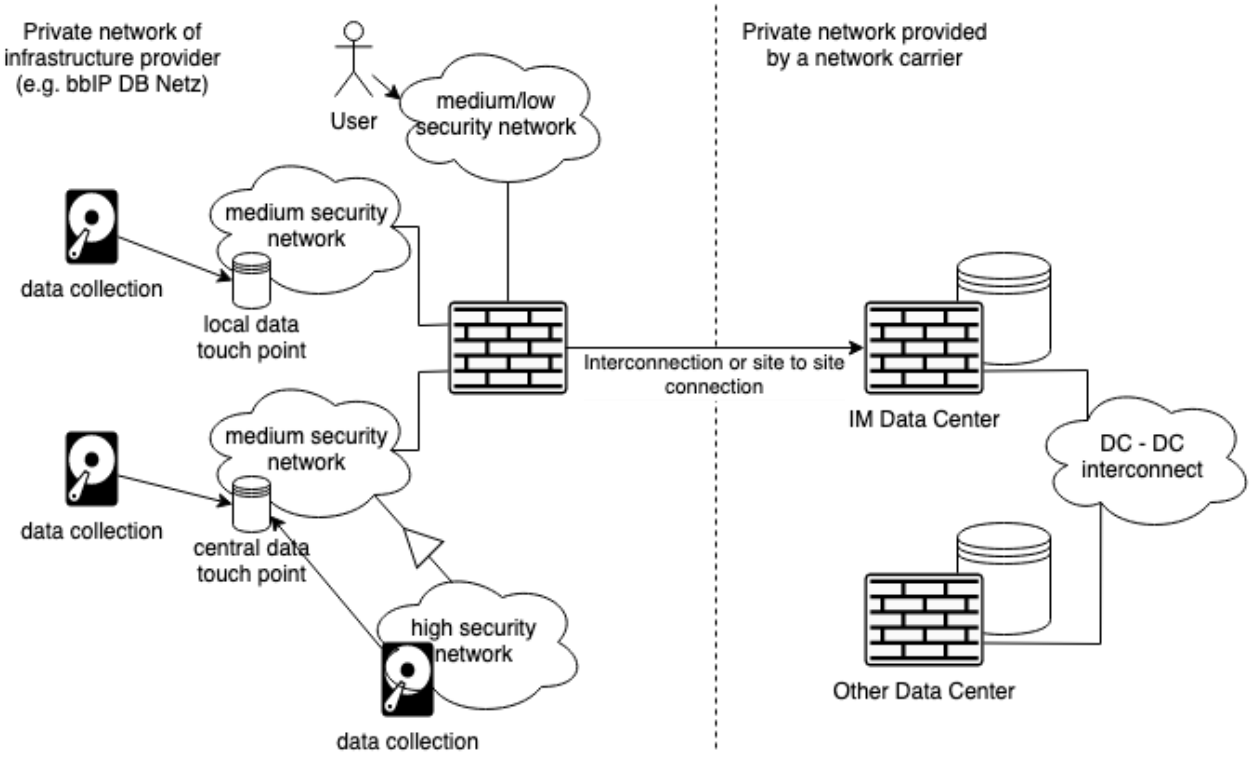

Figure 7: Implementation proposal leveraging private interconnection.

4.2 IMPLEMENTATION ON PUBLIC INTERCONNECTION

Another option for interconnecting with an IM's operational networks is to use public networks, as shown in Figure 8. In this scenario, also user access is facilitated through public networks. This requires additional effort for security, authentication and authorization in the data center as well as the operator's private network in order to protect these from threats originating on the public network

used for data exchange. However, modern networking approaches like SD-WAN may turn this into a viable implementation option.

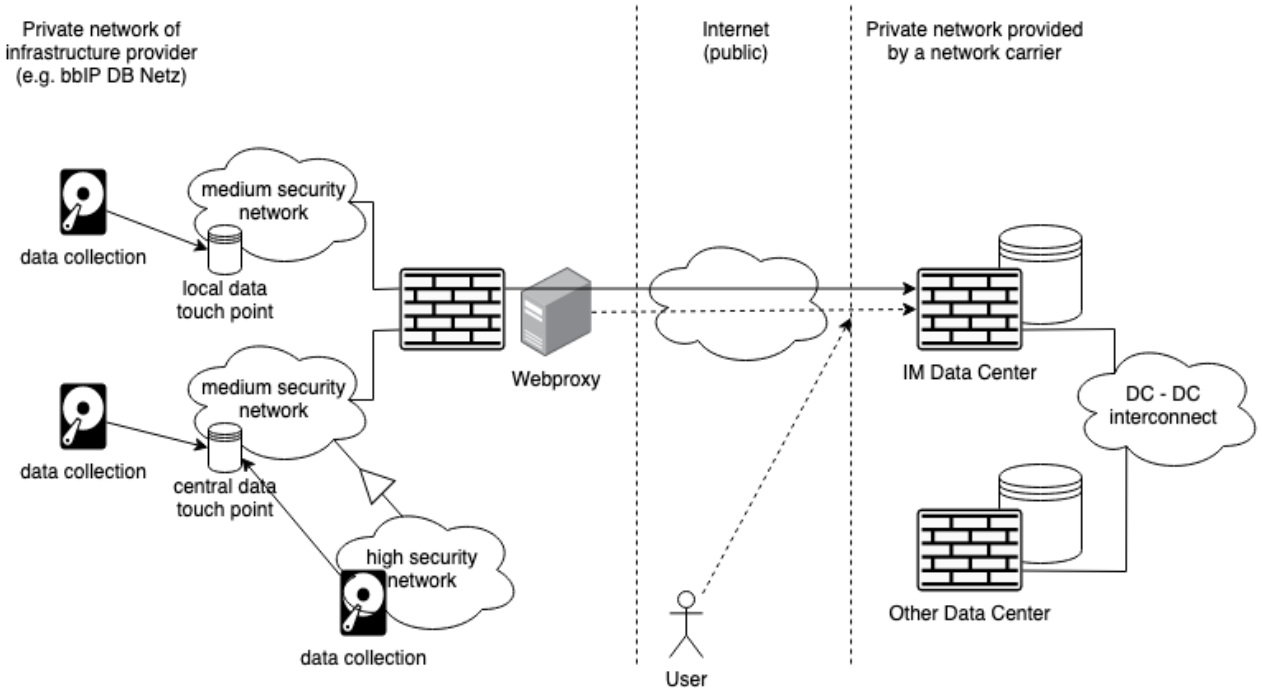

Figure 8: Implementation proposal leveraging public interconnection.

5 CONCLUSION AND OUTLOOK

In the first deliverables D 2.1 and D 2.2 of the CEF2 RailDataFactory Work Package 2, the overall architecture of a pan-European Data Factory has been introduced, as well as concepts for security and data management, including the definition of requirements and open points.

Deliverable D 2.3 has now provided an overview over available networks and possible routes to setting up a data collection backbone, as well as a pan-European Data Factory backbone network. In the process, several implementation options for both were explored and presented, including their comparison. One implementation option was singled out as being the most suited for further discussion. Specific topics for further research, such as operational networks, governance and cost aspects, have been identified.

In the coming deliverables D 3.x and D 4.x, a commercial and operational assessment of the pan-European Data Factory and deployment strategies for the pan-European Data Factory will be presented, respectively.

REFERENCES

[1] Shift2Rail program, see https://rail-research.europa.eu/about-shift2rail/

[2] Europe's Rail program, see https://projects.rail-research.europa.eu/

[3] Sensors4Rail project, see "Sensors4Rail tests sensor-based perception systems in rail operations for the first time," Digitale Schiene Deutschland, 2021. [Online]. Available: https://digitale-schiene-deutschland.de/en/Sensors4Rail

[4] CEF2 RailDataFactory Deliverable 1, "Data Factory Concept, Use Cases and Requirements", Version 1.1, May 2023. [Online]. Available: https://digitale-schiene-deutschland.de/Downloads/2023-04-24_RailDataFactory_CEFII_Deliverable1_published.pdf

[5] Shift2Rail TAURO project, Horizon 2020 GA 101014984, see https://projects.shift2rail.org/s2r_ipx_n.aspx?p=tauro

[6] R2DATO project, see https://projects.rail-research.europa.eu/eurail-fp2/

[7] P. Neumaier, "Data Factory - "Data Production" for the training of AI software," Digitale Schiene Deutschland, 2022. [Online]. Available: https://digitale-schiene-deutschland.de/news/en/Data-Factory

[8] E. Seidler, B. Reichert and C. Kittler, „Das bahnbetriebliche IP-Netz alsSchlüssel für die Digitalisierung der Schiene", SIGNAL+DRAHT 12/2021(https://bit.ly/3OIKcjc)

[9] P. Marsch, R. Fritzsche, B. Holfeld and F.-C. Kuo, "5G for the digital rail system of the future – the prospects for FRMCS", Signal & Draht, March 2022. [Online] Available: Signal-Draht (114) 03_22_5G für das digitalisierte Bahnsystem der Zukunft.pdf (digitale-schiene-deutschland.de)

[10] "DIRECTIVE (EU) 2022/2555 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL", 14 December 2022, https://eur-lex.europa.eu/eli/dir/2022/2555/oj

[11] GAIA-X. [Online] Available: GAIA-X - Home (data-infrastructure.eu)

[12] Catena-X. [Online] Available: Home | Catena-X

[13] IEC TS 62443-1-1:2009 "Industrial communication networks - Network and system security - Part 1-1: Terminology, concepts and models", July 2009. [Onliine] Available: https://webstore.iec.ch/publication/7029